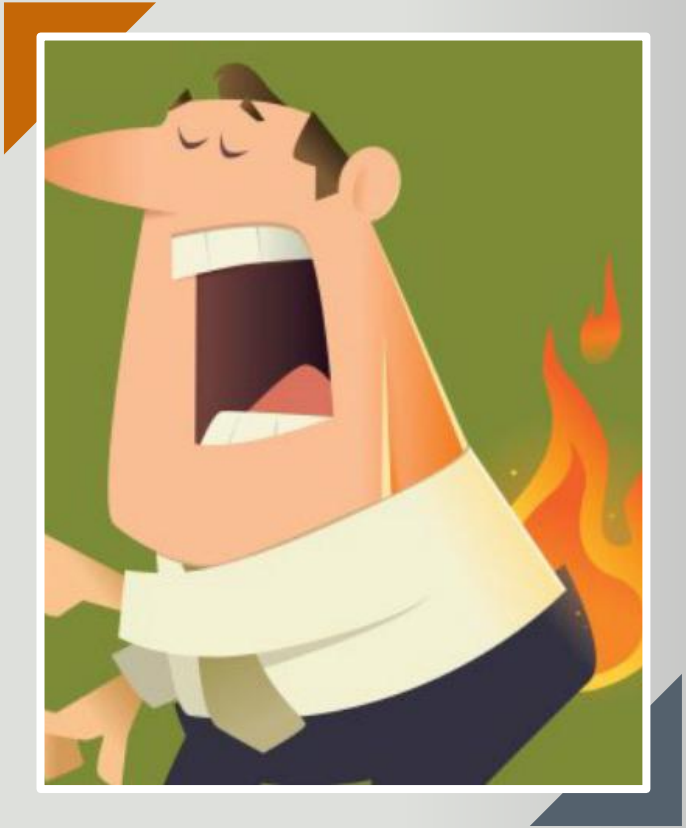


Liar Liar Pants on Fire

A New Benchmark Dataset for Fake News Detection
by
William Yang Wang

Introduction

- Automatic fake news detection is a challenging problem
- Poses threats to the integrity of journalism
- Develops fear and panic situation among people
- Eg
    - Current situation due to COVID -19
    - A man walked into a restaurant with AR -15 rifle upon hearing news of young children being used as sex slaves

I just heard first hand that a doctor who had Corona virus recovered in double quick time. He inhaled Steam just as we normally would in a bowl with towel

• MISLEADING

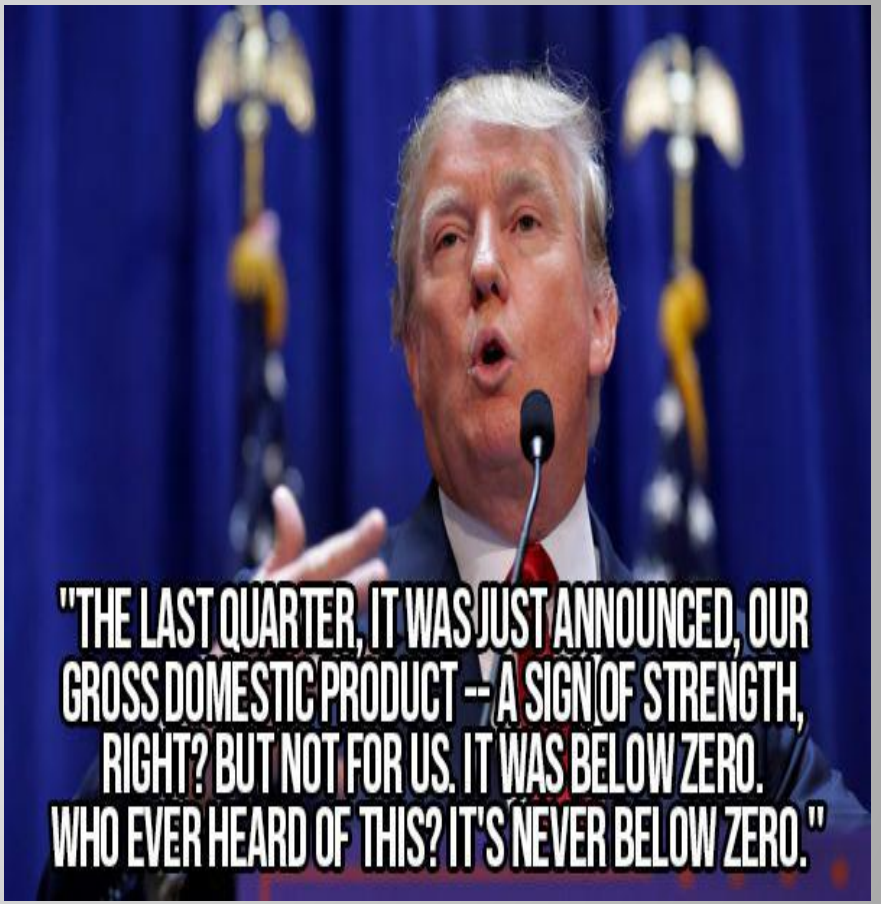
Steaming raises the temperature of lungs, throat and mouth so that if the virus is already there it gets inactivate due to high temperature.

Please also pass this information for the benefit of others.

Like Share

293

33,209 shares



- Automatic fake news detection is a challenging problem similar to deception detection
- Vlachos and Riedel (2014) were the first to release a public fake news detection and fact-checking dataset
- This dataset included only 221 statements which is not feasible for machine learning based assessments
- Crowdsourcing : an important approach to create labeled training dataset
- Results are suboptimal as the algorithm is tested on real world review datasets

Problem Description

- **The main problem in this area is the lack of manually labeled fake news dataset**
- **Other problems include developing machine learning algorithms to improve the accuracy**
- **Crowdsourced datasets are not suitable for fake statements detection because positive training data are collected from a simulated environment**
- **Some of the earlier released datasets are:**
    - **Vlachos and Riedel (2014) were the first to construct fake news and fact-checking datasets.They obtained 221 statements from CHANNEL 42 and POLITIFACT.COM**

- **Ferreira and Vlachos (2016) have released the Emergent dataset, which includes 300 labeled rumors from PolitiFact.**
- **Impractical to use these datasets for developing machine learning algorithms for fake news detection.**

LIAR Dataset

- Significant to introduce larger dataset to facilitate the development of computational approaches fake news detection and automatic fact detection
- The LIAR dataset includes 12.8K human labeled short statements from POLITIFACT.COM's API, and each statement is evaluated by a POLITIFACT.COM editor for its truthfulness.
- We consider six fine-grained labels for the truthfulness ratings: pants-fire, false, barely true, half-true, mostly-true, and true

- **The speakers in the LIAR dataset include a mix of democrats and republicans, as well as a significant amount of posts from online social media.**
- **These statements are sampled from various of contexts/venues, and the top categories include news releases, TV/radio interviews, campaign speeches, TV ads, tweets, debates, Facebook posts, etc**
- **There is also a diverse set of subjects discussed by the speakers**

Contents of Liar Dataset

- ID of the statement
- Label
- Statement
- Subject(s)
- Speaker
- Speaker’s job title
- State info
- Party
- Total credit history account
- True Counts and False Counts
- Pants on Fire Counts
- The Context (Venue / Location of the speech or statement)

LIAR Dataset Statistics

Dataset Statistics	
Training set size	10,269
Validation set size	1,284
Testing set size	1,283
Avg. statement length (tokens)	17.9
Top-3 Speaker Affiliations	
Democrats	4,150
Republicans	5,687
None (e.g., FB posts)	2,185

Table 1: The LIAR dataset statistics.

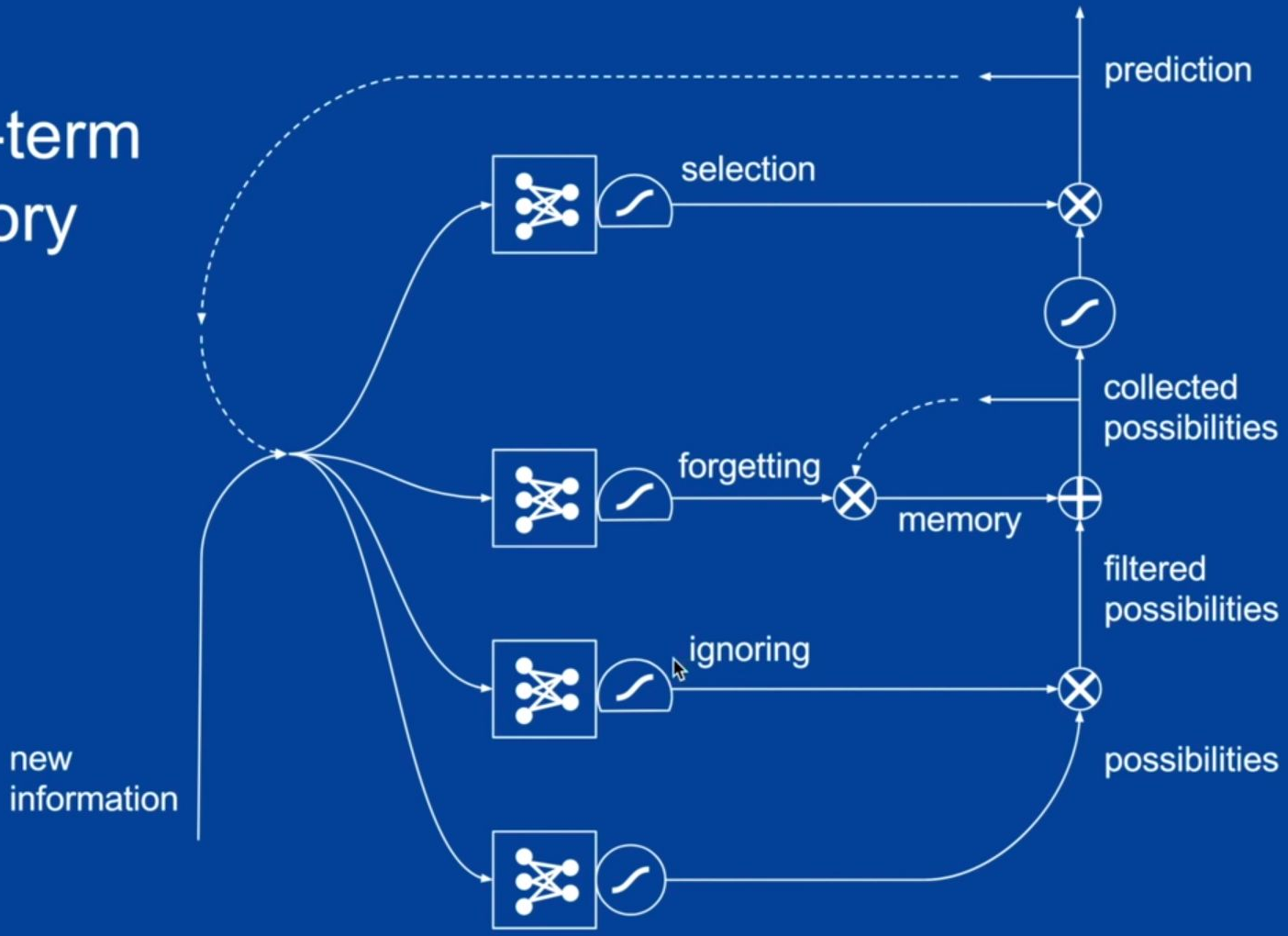
Proposed Solutions on this Dataset

- Some baseline models were used to test on text-only models which includes some popular machine learning models like SVM,Logistic Regression, Bi-LSTM and a CNN.
- The Neural Network to test text +meta-data is a Hybrid CNN, based on a CNN model proposed by another paper (Yoon Kim,2014).
- Accuracy is chosen as the evaluation metric.

Experimental Configuration

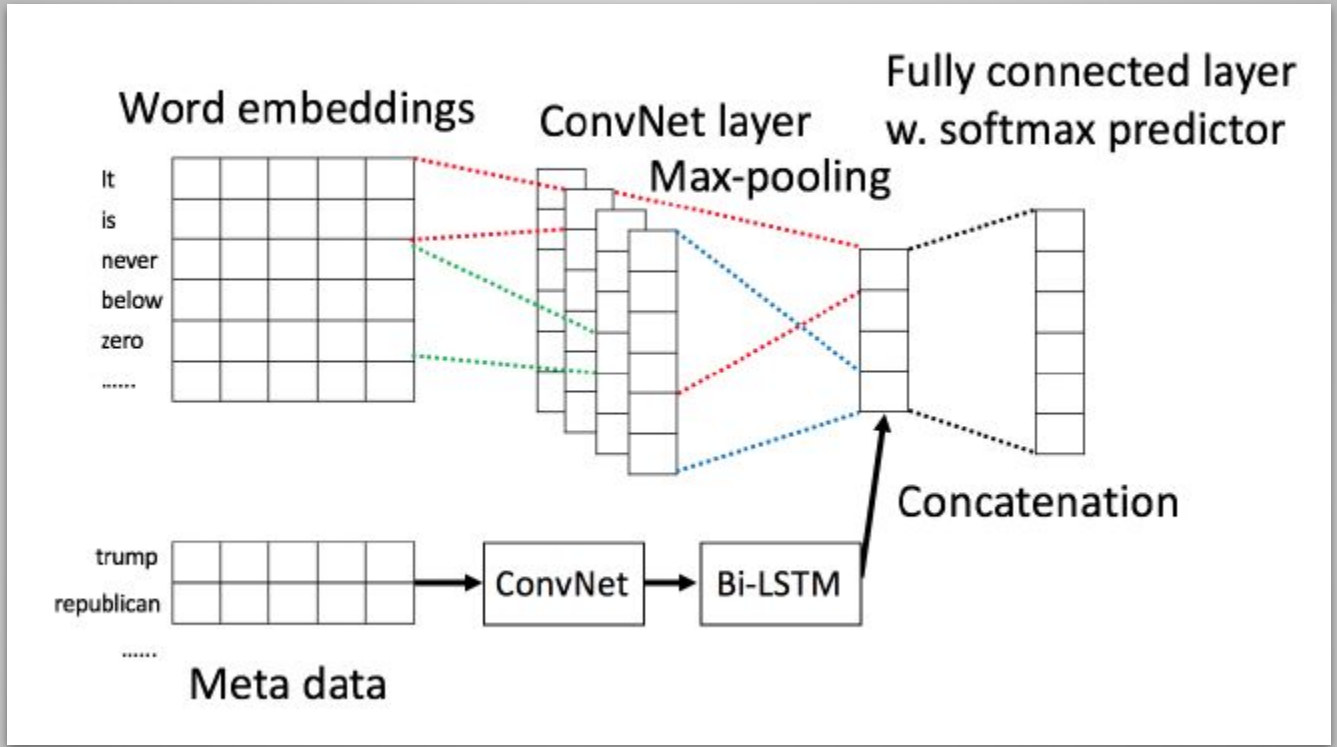
- Grid Search was used to tune Logistic Regression and SVM. These model work well for short text classification problems.
- For the baseline CNN model, 128 filters of size (2,3,4) are used for each ConvNet Layer. The batch size for stochastic gradient descent optimization is set to 64, and the learning process involves 10 passes over the training data for text model.
- For the hybrid model, we use 3 and 8 as filter sizes, and the number of filters was set to 10. We considered 0.5 and 0.8 as dropout probabilities to prevent overfitting.
- The Hybrid CNN model requires 5 training epochs.

long short-term memory



Hybrid CNN

- Randomly initialize a matrix of embedding vectors to encode the metadata embeddings.
- Meta-data vectors are passed through a convolutional layer. Max-pooling is applied , then they are passed to a Bi-directional LSTM layer.
- Text representations are passed through a ConvNet and MaxPool Layer.
- Concatenate the max-pooled text representations with the meta-data representations from the bi-directional LSTM.
- Finally the concatenated output is fed to the Fully Connected Layer with a softmax activation function to generate the final prediction.



The proposed hybrid Convolutional Neural Networks framework for integrating text and meta-data

Results

- Standard text classifier (SVMs and LR models) obtained significant improvements.
- Due to overfitting, the Bi-LSTMs did not perform well.
- The CNNs outperformed all models, resulting in an accuracy of 0.270 on the heldout test set.
- When considering all meta-data and text, the model achieved the best result on the test data.

Models	Valid.	Test
Majority	0.204	0.208
SVMs	0.258	0.255
Logistic Regress0ion	0.257	0.247
Bi-LSTMs	0.223	0.233
CNNs	0.260	0.270
Hybrid CNNs		
Text + Subject	0.263	0.235
Text + Speaker	**0.277**	0.248
Text + Job	0.270	0.258
Text + State	0.246	0.256
Text + Party	0.259	0.248
Text + Context	0.251	0.243
Text + History	0.246	0.241
Text + All	0.247	**0.274**

The evaluation results on the LIAR dataset

Conclusion

- This Paper shows that when meta-data is combined with text, significant improvements can be achieved for fine-grained fake news detection.
- The LIAR dataset is a magnitude larger with real-world short statements from various contexts with diverse speakers.
- This corpus can also be used for stance classification, argument mining, topic modeling, rumor detection, and political NLP research.
- Given the detailed analysis report and links to source documents in this dataset, it is also possible to explore the task of automatic fact-checking over knowledge base.